Explaining the Black-box Smoothly- A Counterfactual Approach

Year: 2021 | Citations: 122 | Authors: Junyu Chen, Yong Du, Yufan He, W. Paul Segars, Ye Li

---

Abstract

We propose a BlackBox Counterfactual Explainer, designed to explain image classification models for medical applications. Classical approaches (e.g., , saliency maps) that assess feature importance do not explain how imaging features in important anatomical regions are relevant to the classification decision. Such reasoning is crucial for transparent decision-making in healthcare applications. Our framework explains the decision for a target class by gradually exaggerating the semantic effect of the class in a query image. We adopted a Generative Adversarial Network (GAN) to generate a progressive set of perturbations to a query image, such that the classification decision changes from its original class to its negation. Our proposed loss function preserves essential details (e.g., support devices) in the generated images. We used counterfactual explanations from our framework to audit a classifier trained on a chest X-ray dataset with multiple labels. Clinical evaluation of model explanations is a challenging task. We proposed clinically-relevant quantitative metrics such as cardiothoracic ratio and the score of a healthy costophrenic recess to evaluate our explanations. We used these metrics to quantify the counterfactual changes between the populations with negative and positive decisions for a diagnosis by the given classifier. We conducted a human-grounded experiment with diagnostic radiology residents to compare different styles of explanations (no explanation, saliency map, cycleGAN explanation, and our counterfactual explanation) by evaluating different aspects of explanations: (1) understandability, (2) classifier's decision justification, (3) visual quality, (d) identity preservation, and (5) overall helpfulness of an explanation to the users. Our results show that our counterfactual explanation was the only explanation method that significantly improved the users' understanding of the classifier's decision compared to the no-explanation baseline. Our metrics established a benchmark for evaluating model explanation methods in medical images. Our explanations revealed that the classifier relied on clinically relevant radiographic features for its diagnostic decisions, thus making its decision-making process more transparent to the end-user.